Aix-Marseille Université **M1 MAS**

Examen de statistique exploratoire du 16 décembre 2020

les documents sous forme papier autorisés sont les polys, feuilles de TP, notes manuscrites,

sont interdits: calculatrices, téléphones, ordinateurs portables

---

On considère des données concernant $p = 14$ caractéristiques physiques de $n = 250$ hommes. Les données sont dans un tableau de n lignes et p colonnes, qui a été importé sous SAS. Les colonnes du tableau correspondent aux 14 variables suivantes (entre parenthèse leur nom donné dans la table SAS):

- le pourcentage de graisse corporelle (PctGC),
- l'âge (Age) en années,
- la taille (Hauteur) en centimètres,
- le poids (Poidskg) en kilogrammes,
- la dimension du cou, de la poitrine, de l'abdomen, de la hanche, de la cuisse, du genou, de la cheville, du biceps, de l'avant bras, du poignet

  (noms SAS: Cou, Poitrine, Abdomen, Hanche, Cuisse, Genou, Cheville, Biceps, AvtBras, Poignet)

  toutes en centimètres.

Remarque 1: l'identité des individus n'a pas été collectée, ils sont juste repérés par une étiquette de ligne allant entre 1 *et* 250

Remarque 2: le jeu de données est quasiment identique à celui qui sera utilisé dans votre examen de modèles linéaires.

Univarié, bivarié

1. On a fait le boxplot de la variable *poidskg* (Figure 1). Commenter l'allure de ce boxplot. (deux ou trois phrases suffisent) A noter: les nombres à côté des points en haut du graphique sont les étiquettes des individus concernés.

2. On a calculé le coefficient de corrélation linéaire entre la dimension de l'abdomen et le pourcentage de graisse corporelle, on obtient: 0.81343.

   En général quelles sont les valeurs possibles d'un coefficient de corrélation linéaire, et par conséquent comment qualifieriez vous la valeur ici présente? Qu'est ce que cela signifie concrètement pour les individus? Aurait on pu s'attendre à ce type de relation? (une phrase pour chaque réponse suffit)

ACP

On fait maintenant une ACP normée sur les 14 variables.

1. Rappelez pourquoi l'inertie totale est $I = 14$ (une phrase suffit)

2. Combien d'axes principaux faudrait il retenir, d'après la règle de Kaiser et les annexes (une phrase suffit)

3. D'après le cercle des corrélations (donné en Figure 2), comment caractériser en fonction des variables initiales les individus qui sont mis en opposition sur l'axe 1, c'est à dire ceux à "gauche" contre ceux à "droite" dans le 1er plan principal (donné aussi en Figure 3)?

4. On mesure maintenant toutes les variables de dimensions en millimètres ou lieu des centimètres. On relance l'ACP normée. Les résultats changeront ils par rapport à l'ACP déjà faite? Justifier en une ou deux phrases.

Classification

On fait maintenant une classification sur les 14 variables centrées et réduites.

1. En général en classification, pour un nombre de classes donné, que vaut il mieux avoir: beaucoup d'inertie interclasse ou beaucoup d'inertie intraclasse? (expliquer en deux ou trois phrases)

2. On calcule le critère "semi partial R square" *sprs* pour chaque étape de l'algorithme. On l'a représenté en Figure 4 pour les dernières étapes: en ordonnée le *sprs* obtenu à l'issue de l'étape de l'algorithme qui conduit au nombre de classes indiqué en abscisse. D'après ce graphique, quel nombre de classes fixeriez vous pour la partition? (une phrase suffit)

3. Dans la suite on effectue la classification en choisissant 2 comme nombre total de classes (attention ce n'est pas forcément le choix dicté par le critère *sprs* vu à la question précédente). On calcule la moyenne de chaque variable initiale dans la classe 1, dans la classe 2, et en général (voir Table 2). On calcule aussi la "valeur test" de chaque variable initiale dans la classe 1 et dans la classe 2. Comme vous le voyez dans la Table 3, dans ce cas particulier, les valeurs tests au sein de la classe 1 sont les opposées de celles de la classe 2 (pas besoin de le démontrer).

   D'après ces valeurs tests, quelle interprétation des classes proposez vous?

4. A votre avis, dans laquelle des 2 classes devrait se trouver le 39ème individu de la table? (une phrase suffit)

ANNEXES

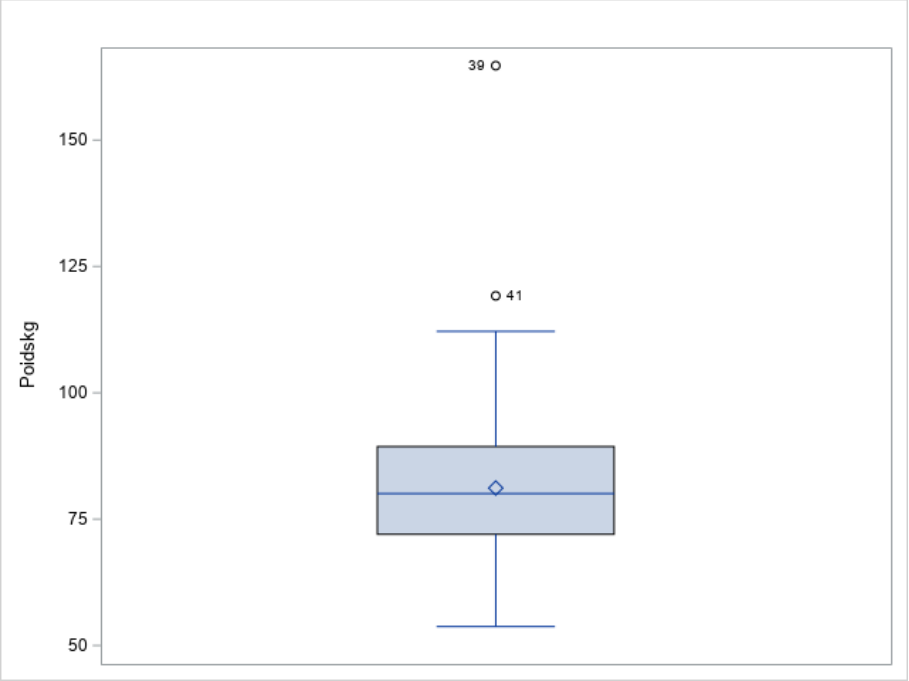

Figure 1: boxplot de la variable *poidskg*

k	lambda	pctvar	cumpct
1	8.46	0.60	0.60
2	1.58	0.11	0.72
3	1.04	0.07	0.79
4	0.67	0.05	0.84
5	0.63	0.04	0.88
6	0.41	0.03	0.91
7	0.30	0.02	0.94
8	0.26	0.02	0.95
9	0.20	0.01	0.97
10	0.18	0.01	0.98
11	0.13	0.01	0.99
12	0.08	0.01	1.00
13	0.04	0.00	1.00
14	0.02	0.00	1.00

Table 1: valeurs propres de l'ACP (1ère colonne), proportion d'inertie expliquée par axe et proportions d'inerties cumulées (2ème et 3ème colonnes)

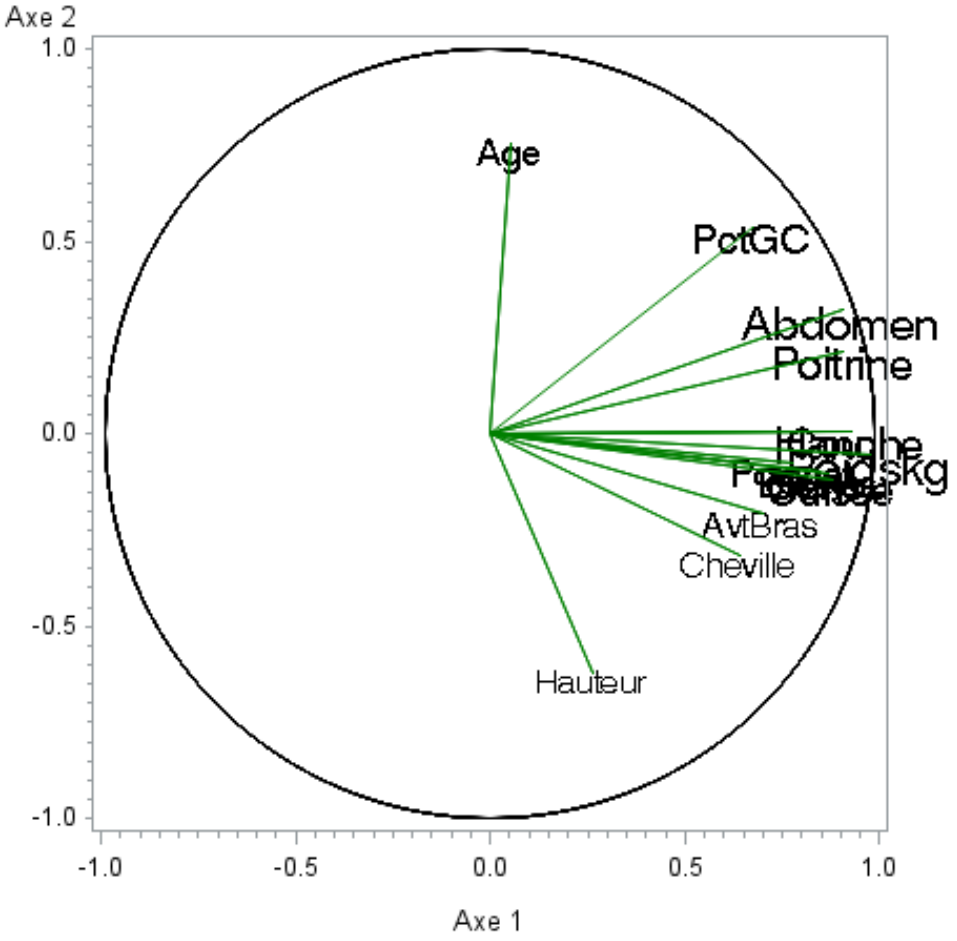

Figure 2: carte des variables dans la plan factoriel 1-2

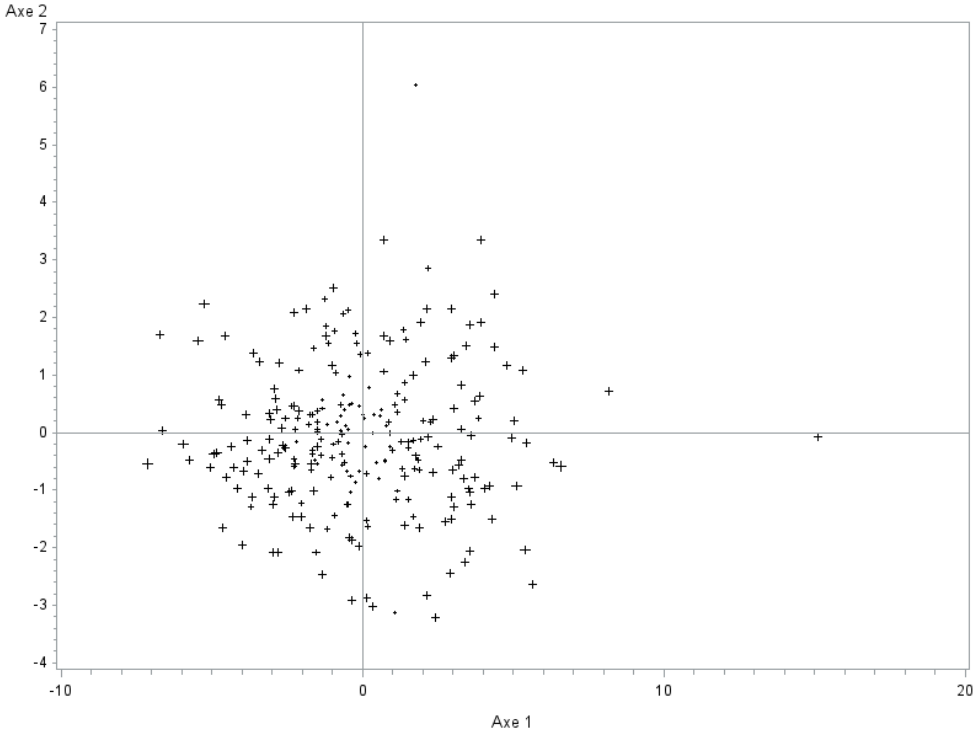

Figure 3: carte des individus dans la plan principal 1-2

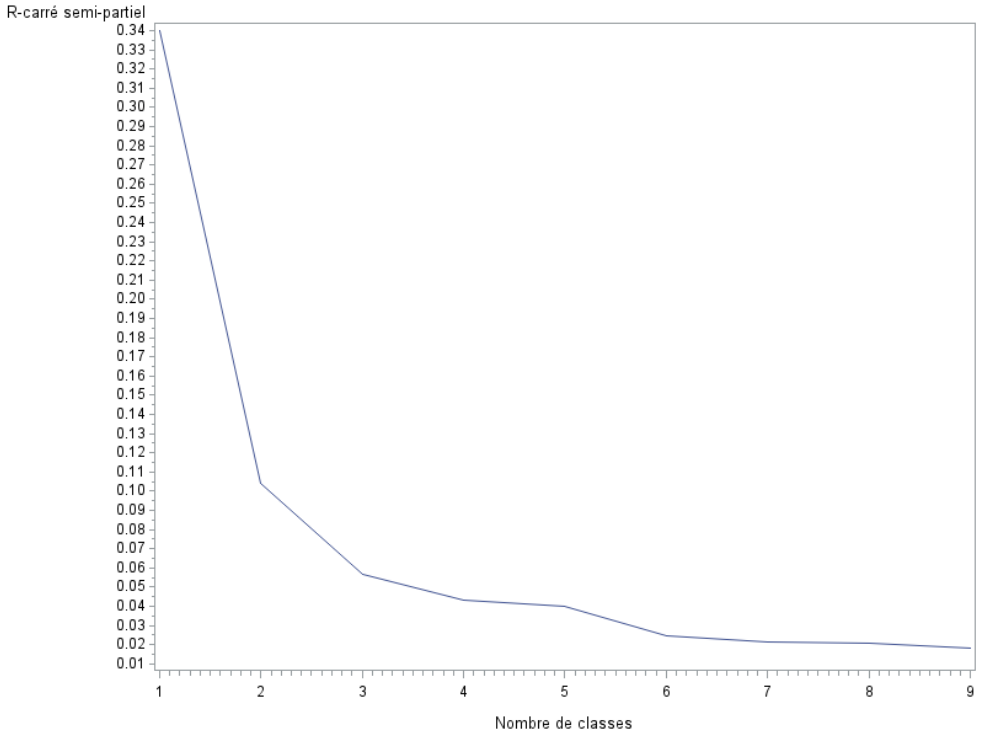

Figure 4: valeurs du semi partial R square pour les dernières étapes de l'algorithme

	classes		Ensemble
	1	2	
	moyenne	moyenne	moyenne
PctGC	22.20	15.69	19.15
Age	44.73	45.06	44.88
Hauteur	180.22	175.86	178.18
Poidskg	90.03	71.09	81.16
Cou	39.40	36.39	37.99
Poitrine	106.07	94.87	100.82
Abdomen	98.73	85.55	92.56
Hanche	104.19	95.04	99.90
Cuisse	62.46	55.93	59.41
Genou	40.04	36.94	38.59
Cheville	23.98	22.10	23.10
Biceps	34.22	30.06	32.27
AvtBras	29.90	27.26	28.66
Poignet	18.80	17.58	18.23

Table 2: moyennes des variables initiales au sein de la classe 1 (1ère colonne), au sein de la classe 2 (2ème colonne), et en général (3ème colonne)

	valeur test	
	Classes	
	1	2
PctGC	6.15	-6.15
Age	-0.21	0.21
Hauteur	3.70	-3.70
Poidskg	11.23	-11.23
Cou	9.80	-9.80
Poitrine	10.50	-10.50
Abdomen	9.66	-9.66
Hanche	10.09	-10.09
Cuisse	9.83	-9.83
Genou	10.16	-10.16
Cheville	8.77	-8.77
Biceps	10.89	-10.89
AvtBras	10.37	-10.37
Poignet	10.35	-10.35

Table 3: valeurs tests des variables initiales pour la classe 1 (1ère colonne), et pour la classe 2 (2ème colonne)